Argonne NATIONAL LABORATORY

... for a brighter future

ALCF

Argonne Leadership Computing Facility

U.S. Department of Energy

UChicago ▶ Argonne LLC

Office of Science U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

Introduction to BGL

Argonne Leadership Computing Facility

Susan Coghlan, Deputy Director
Argonne National Laboratory and University of Chicago

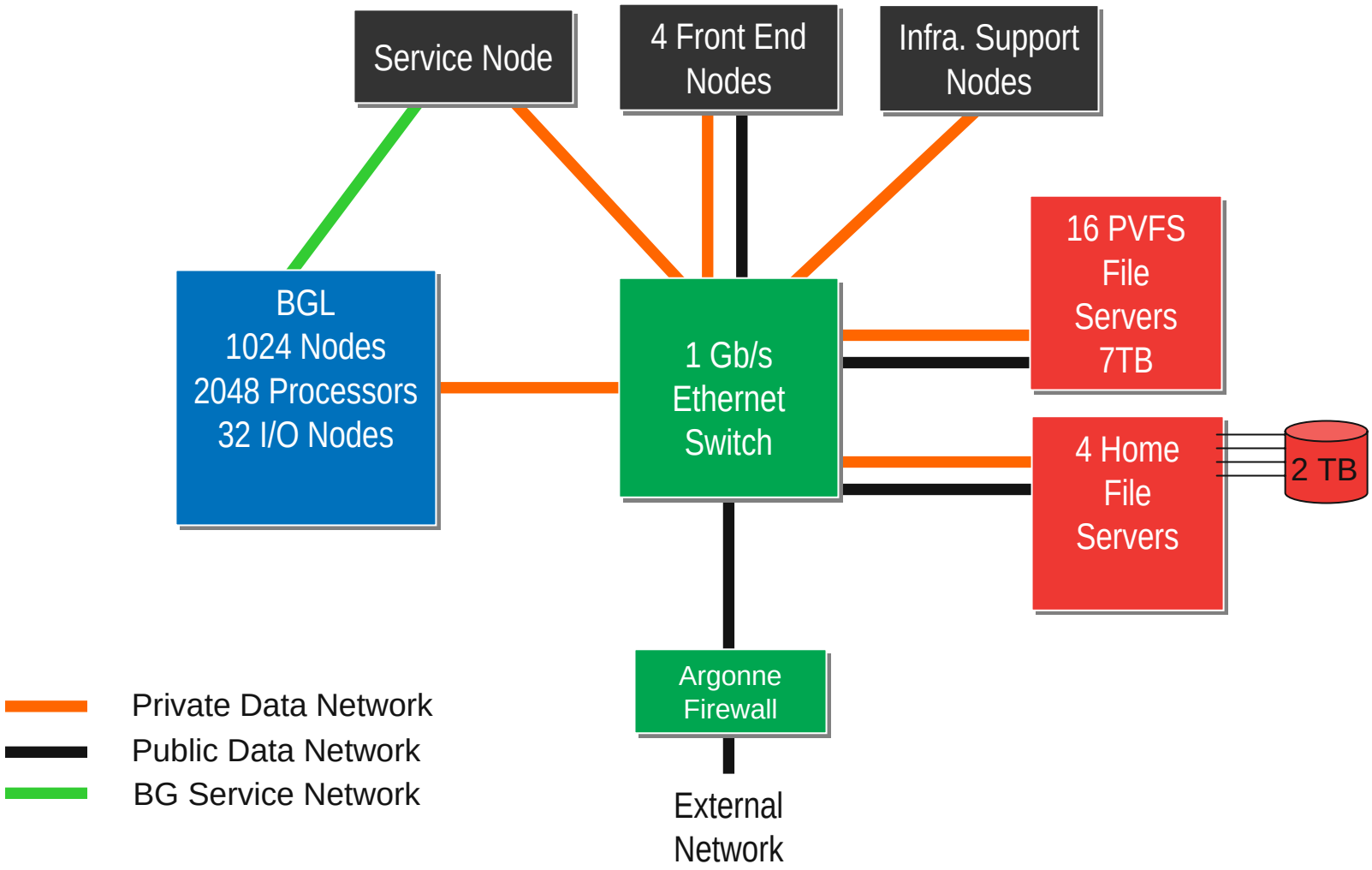
February 7, 2007

Overview

- What this talk covers:
  - Layout of the BGL system
  - Getting onto the system and storing your code and data
  - Compiling a simple program
  - Submitting a job
  - Figuring out simple problems

- What this talk does *not* cover:
  - Compiling complex programs
  - Using a non-default network layout
  - Doing high performance I/O
  - Deep debugging

BGL System Architecture



Configuration Details

- Login servers [4]
  - Compile and submit jobs
  - bgl.mcs.anl.gov -> 2 servers DNS round-robin
  - login[1-4].bgl.mcs.anl.gov
- Service Node [1]
  - All jobs are started from the service node
  - It must be able to see the executable and starting directory (cwd)
  - Users have restricted shells on this server
- I/O nodes [32]
  - 1/32 IO node/compute node ratio
  - Computes are mapped to a specific IO node
  - ssh access allowed thru ZeptoOS kernel
- Compute nodes [1024]
  - No direct access
- Storage servers [20]
  - Admin and developer access only

I/O on BGL

- Home directory
  - /bgl/home1/<username> (Aliased to /home/<username>)
  - Visible on: login servers, I/O nodes, computes, Service Node
  - Limited space (please watch usage)
  - Backed up nightly
  - Not a good idea to use for large quantity of accesses during job runs
- Local disk
  - /sandbox - **only** on the login servers, do not use for actual jobs
  - Scratch space - **not backed up!**
  - No local disk available on computes or I/O nodes
- Data
  - /pvfs/<username>
  - Visible on: login servers, I/O nodes, computes
  - Not visible to Service Node, so, no exec and no stderr/stdout files
  - **Not backed up!**

Building Executables

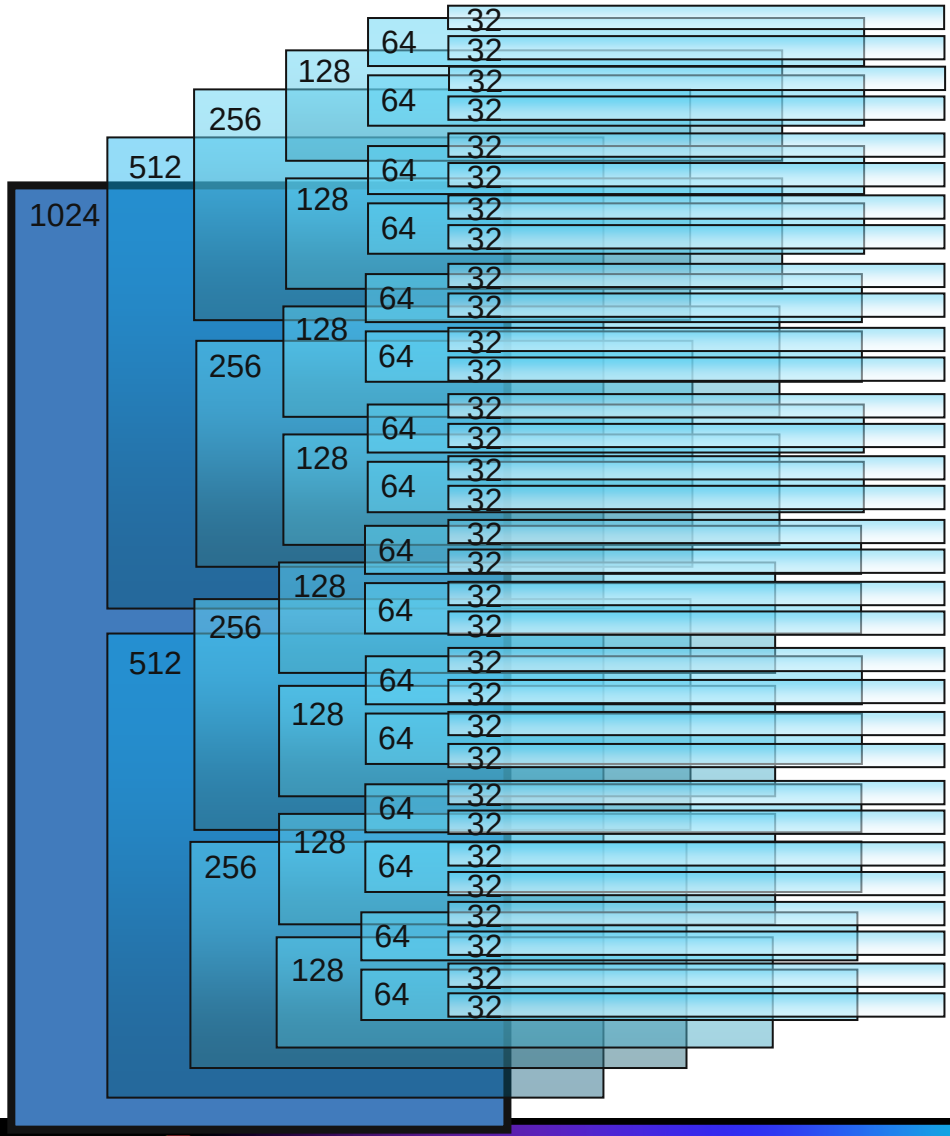
- MPI wrappers (easiest): mpi<language>.<compiler>

mpicc.ibm	mpicxx.ibm	mpif77.ibm	mpif90.ibm
mpicc.gnu	mpicxx.gnu	mpif77.gnu	

  ex: mpicc.ibm -o HelloWorld.rts HelloWorld.c

- Be careful about mpicc vs mpicc.ibm
  - mpicc, mpicxx, mpif77 are IBM shipped wrappers that use the gnu compilers and have some problems
- Direct compiler and library linking also possible:
  - *Ex: /opt/ibmcmp/xlf/9.1/bin/blrts_xlf*
  - details in the **Hints & Tips** handout
- Optimizations
  - details in the **Hints & Tips** handout
  - advisors will assist with optimizations beyond the standard set
- Note: IBM compilers are generally recommended over GNU ones

BGL Partitions ("blocks")



- 1 I/O node for each 32 compute nodes, hardwired to specific set of 32
  - *Minimum partition size of 32 nodes*
- Partition sizes: 32, 64, 128, 256, 512, 1024
  - *Any partition < 512 nodes will get a mesh network layout and not a torus.*
  - *Any partition <512 nodes will get a non-optimal I/O tree network.*
  - *Do not do performance testing on <512 nodes*
- Smaller partitions are enclosed inside of larger ones
  - *Not all partitions are available at all times*
  - *Once a job is running on one of the smaller partitions, no jobs can run on the enclosing larger partitions*
- Configuration changes frequently
  - ***partlist** shows partition state*
- Processes are spread out in a pre-defined mapping (XYZT), alternate and sophisticated mappings are possible

Resource Mgr and Job Scheduler

- Cobalt - locally developed
- Standard commands, but prefaced with a ‘c’:
  - cqsub: submit jobs
  - cqstat: check job status
  - cqdel: delete jobs
- Queues (FIFO based, with some exceptions)
  - **default** - no need to specify
  - **short** - only jobs <= 30 minutes long, #nodes varies
  - **incite** - special queue for incite reservations
  - **incite-workshop** - use for all of your jobs today and tomorrow
  - Special purpose queues, generally for specific people and needs
- Reservations
  - Required for anything larger than 512 nodes
  - Production runs will need to be under reservations
  - Email reservation requests to ***support@bgl.mcs.anl.gov***
- Standing reservations
  - Preventative maintenance reservation: Each Monday at 5pm
  - Big Run Day: Each Tuesday 10-5 (depending on requests)

Submitting Simple Jobs

- *cqsub -q short -t 00:10:00 -n 32 HelloWorld.rts*
  - *Will run on partitions in* short *queue*
  - *Will end after 10 minutes or when the executable exits,* whichever comes first
  - *Will run on 32 nodes, 32 processors*
  - *Output will be stored in <jobid>.output and <jobid>.error*

  ***Warning**: don't request less than 5 minutes*

  ***Note**: always add 5 minutes to the time that you need to allow for booting the partition and other overhead.*

- *cqsub -t 01:00:00 -n 256 -m vn HelloWorld.rts*
  - *Will run on 256 nodes with 512 processors (due to -m vn)*

- '*man cqsub'* for details about possible options

Submitting More Complex Jobs

- *cqsub -t 01:00:00 -n 128 -c 256 -m vn HelloWorld.rts*
  - *Will run on partitions in* default *queue*
  - *Will run on 128 nodes, 256 processors*

  Warning**:* **-m vn *required to get 256 processors

- *cqsub -t 01:00:00 -n 256 -m vn -e BGLMPI_MAPPING=TXYZ HelloWorld.rts*
  - *Will run on 256 nodes with 512 processors (due to -m vn)*
  - *Use -e to pass in environment variables (see the Hints & Tips document for other examples)*

  Note: It is a good idea to use TXYZ mapping for vn mode jobs since the default mapping puts processor 0 and 1 far away from each other rank-wise. This is important if locality is inherent in the code.

Why doesn't my job run?

- Possible causes:
  - Pending reservation
    - *Note: There is 10 mins of dead time prior to reservation start (i.e., if there is 60 mins until the reservation starts, any job longer than 50 mins will not be started).*
  - No partitions available
  - Wrong queue
  - Partitions not freed
- Use '***cqstat***' to see both running and waiting jobs
  - ***cqstat -f*** - show more details (queue, etc.)
  - ***cqstat -fl*** - show even more details (executable name/args, stdin/stdout paths, etc)
  - Status: Q waiting, R running
- ***showres*** - show all defined reservations (pending and not yet deleted)
- ***partlist*** - show online partitions and status (sort of)
- ***bgl-listblocks*** - show partition info
- Sometimes a job disappears from queue but is still holding a partition - '*bgl-listblocks'* can show if a partition is still allocated*, 'bgl-listjobs'* will show jobs that BGL believes are still running.

My job is no longer in the queue, but I don't think it ran successfully…

- First place to look: STDERR file <jobid>.error (default)
  - Sometimes the error messages are obscure - send mail to support.
  - Note: Two job ids - Cobalt and BGL, both are important. This jobid is the Cobalt jobid.
  - STDOUT file is <jobid>.output (default)
  - Some common errors:
    - *"**Killed with signal -9**" usually means the job ran out of time (if you did not cqdel the job). Check the times at the top, bottom of .error file and the requested walltime.*
    - *"**Job deleted because block was deallocated**" often indicates hardware failure. Contact support with location of error file.*
- ***bgl-listevents*** - show events associated with job
  - Lots of options, use ***bgl-listevents -h*** for help with options.
  - If you need help interpreting, or finding any events, send mail to support.

A few other things to check…

- Are there any core files?
  - core.<node#>
  - ascii files, use bgl_stack to decode
  - If you need help interpreting, send mail to support.
- Can you run a simple HelloWorld successfully?
  - If not, have you changed your dot files?
  - Are you forwarding X thru your ssh?
- Are your CWD and executable within your home directory space?
- Use print statements, but be aware that I/O is **very** buffered.
- If all else fails, there is an extremely limited version of gdb
  - You will need to request a partition for direct running of your job (i.e., not thru Cobalt).
  - Send mail to support.

Tools: Debugging & Others

- GDB - method of last resort, you will need to work with support
- Currently two places for tools and apps: /soft/apps and /soft/tools
- Heap/stack memory collision protection/tracking
- Tracing 'exit' and 'abort'
- Libraries:
  - BLAS, LAPack
  - Mass, MassV
  - ESSL - very old version
  - FFTW
  - hdf5, netcdf
  - PETSc
  - Profiling
  - TAU
  - "-qdebug=function_trace"
- Your advisors will provide more details.
- More on profiling later

Help!

- System issues (e.g., jobs not being scheduled, access problems, reservation requests,system not responding, etc.)

  email to support@bgl.mcs.anl.gov

- General BG/L questions and problems (e.g., what optimization flags work best, what libraries are available, how mapping works)

  email to discuss@bgl.mcs.anl.gov

- Resources:
  - BGL Hints & Tips document online:

    bgl.mcs.anl.gov:/software/common/doc/BGL-Hints-Tips.txt
  - BG/L wiki: http://wiki.bgl.mcs.anl.gov
  - BGL web pages: http://bgl.mcs.anl.gov
    - *Good starter page: http://bgl.mcs.anl.gov/Documentation (contains link to online IBM redbooks)*
    - *System Administration (Oct 30, 2006), Hardware Overview and Planning (Aug 11, 2006), Application Development (Jan 18, 2007), Performance Analysis Tools (Jul 18, 2006), Solution Problem Determination (Oct 11, 2006)*